Recommendation system:
Connecting business users with innovative solutions

Candidate: Ivana Nastasic

Master Course: Data Science

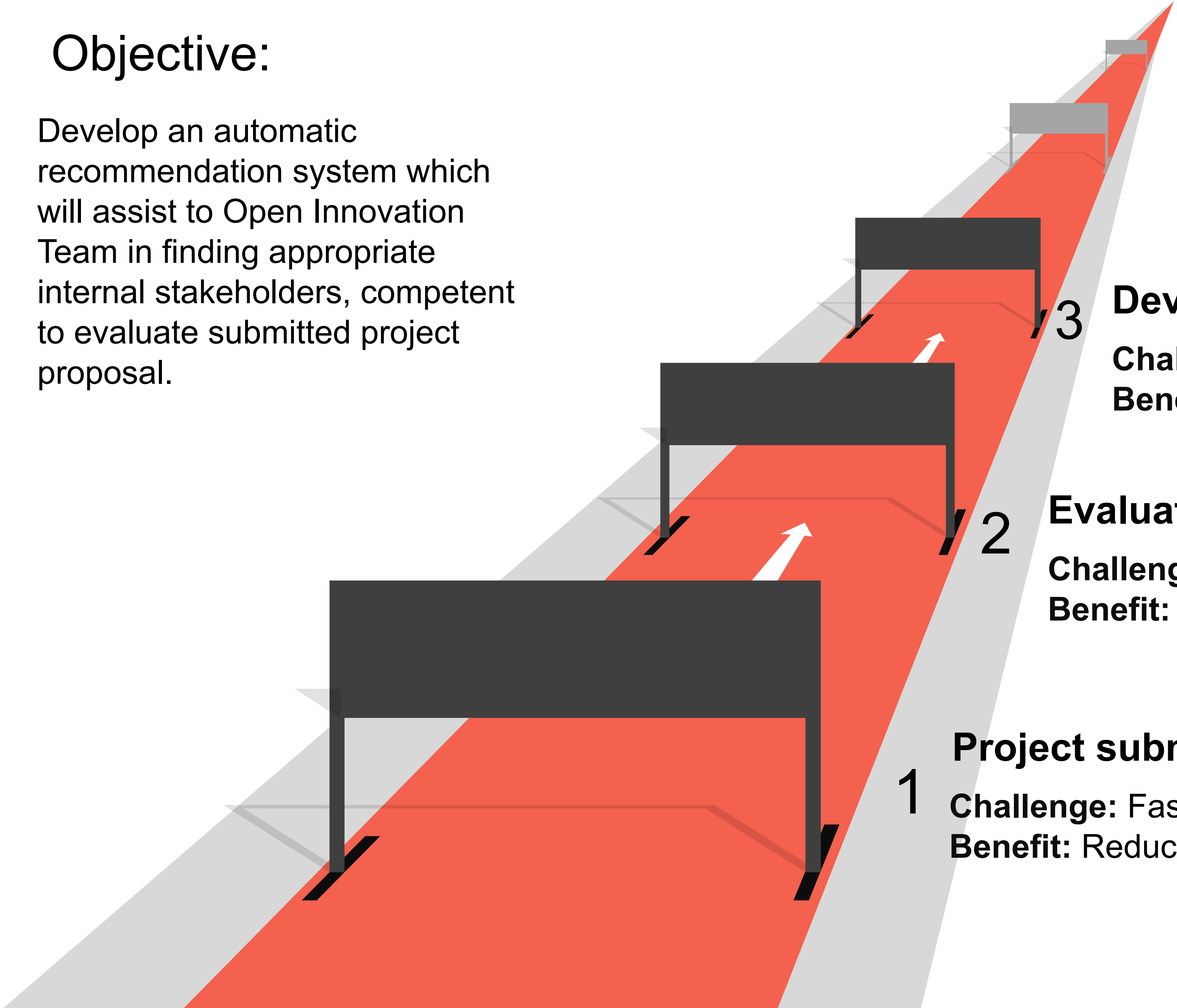
Thesis Advisor: Prof. Pierpaolo Brutti
External Advisor: Marco Piersanti

Sapienza università di Roma - Facoltà di ingegneria dell'informazione informatica e statistica

Introduction

Objective:

Develop an automatic recommendation system which will assist to Open Innovation Team in finding appropriate internal stakeholders, competent to evaluate submitted project proposal.



3 **Development**

Challenge: Time consuming
Benefit: Decrease effort of the Open Innovation Team

2 **Evaluation by a technical panel of experts**

Challenge: Complex internal organization
Benefit: Increase evaluation quality

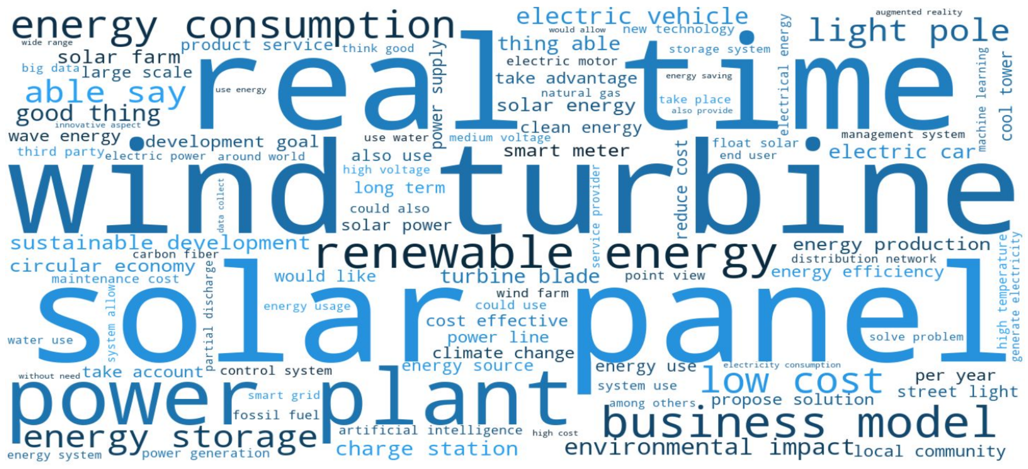
1 **Project submission via ENEL Open Innovability Portal**

Challenge: Fast evolving technology
Benefit: Reduce evaluation feedback time

Datasets

Row data export from The Open Innovability Portal, ~3800 Project Descriptions in several languages.	
Project Proposals dataset example	
SOL-27358	There is a plugin Office called "Dictate" . It can be downloaded from this Microsoft website (dictate.ms) . Using this plugin Office programs (Outlook,Word,Powerpoint...) can write automatically or translate. in another Language. The benefit is that for an Enel employee, is easier and faster, to think a document and to speak to this "digital secretary", then to type on the desk. To be more clearer, please look at this 2 youtube walk-through videos: https://www.youtube.com/watch?v=auF9bvAectU https://www.youtube.com/watch?v=k9gCfEJGj38

Processed dataset based on self presentation of employees on internal e-profile portal, ~ 35000 employees, 18 skill types and 298 skill subtypes. Skills are entered in free text format.	
Employee's skills dataset example	
Employee ID:196 Skill ID: 1131248816-2	Skill type: energy related skills Skill subtype: power plants Skill description: power generation management

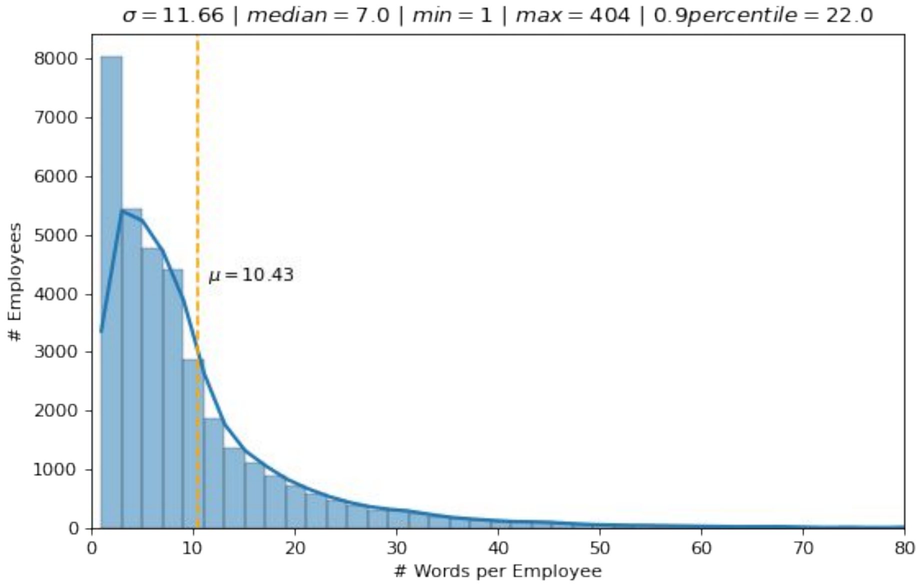


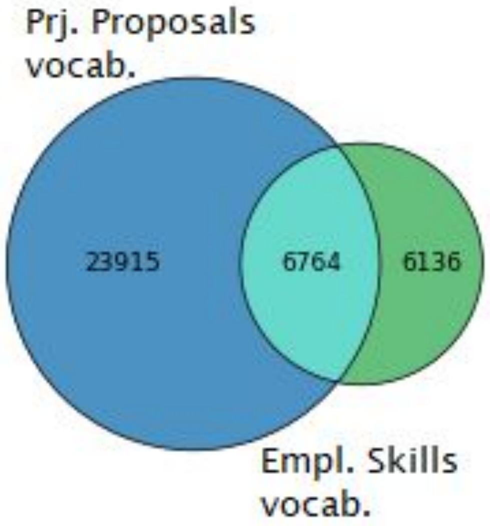
Exploratory Data Analysis Findings

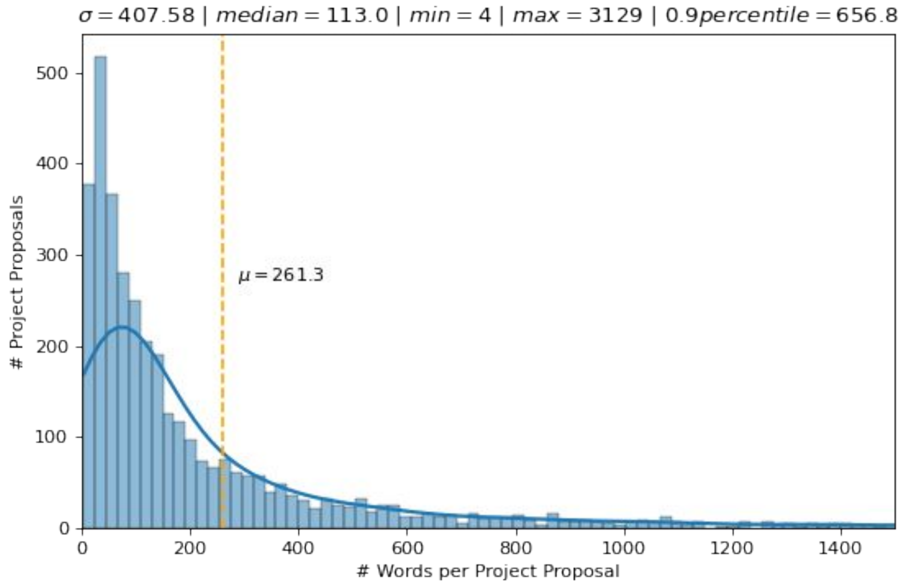
01 Text descriptions of employee's skills are very short.

02 Vocabularies differ significantly in size and content.

03 There is a huge difference in text length between projects and employees descriptions.



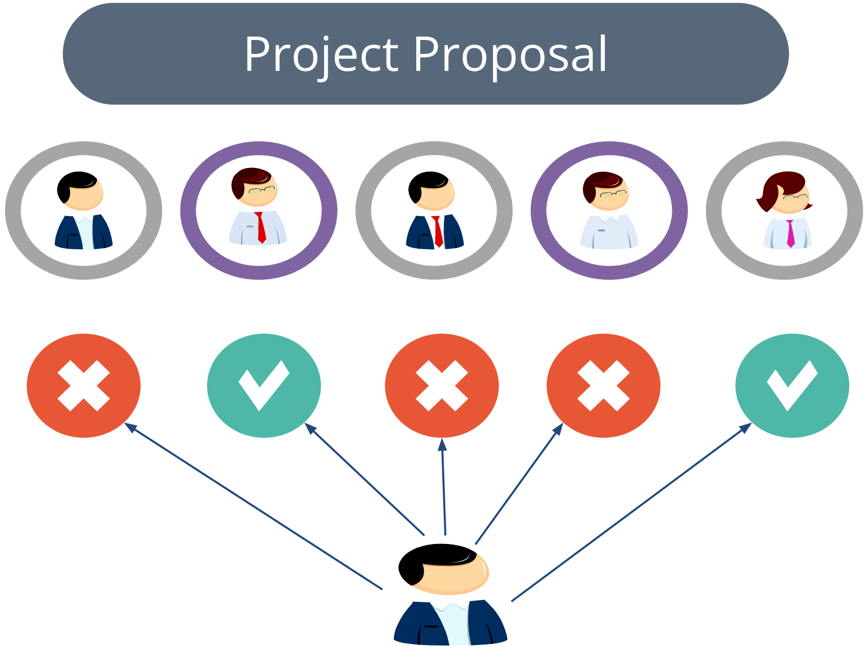




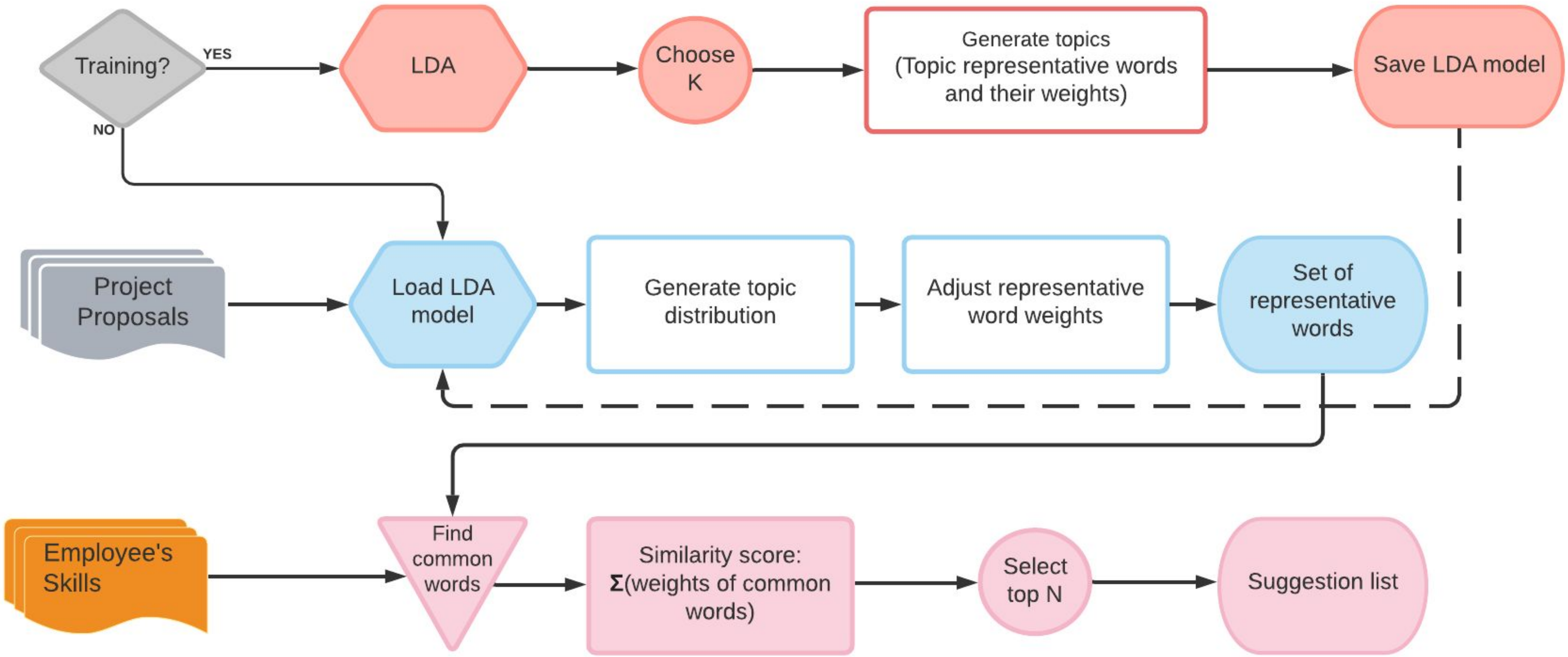
Evaluation

01 User-centric Perceived Recommended Accuracy: % of project proposals with at least 1 relevant suggestion.

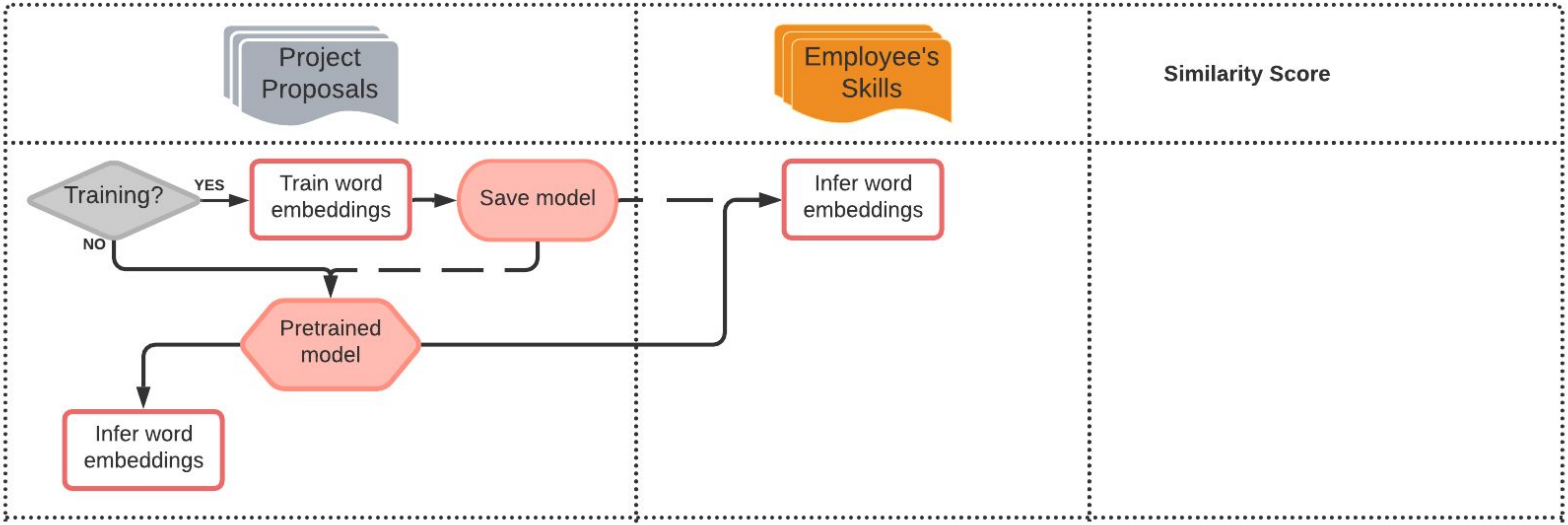
- **20** Project Proposals
- **5** Suggested Employees
- **10** Perceived Relevance
- Open Innovation Team
- **3** Model setups



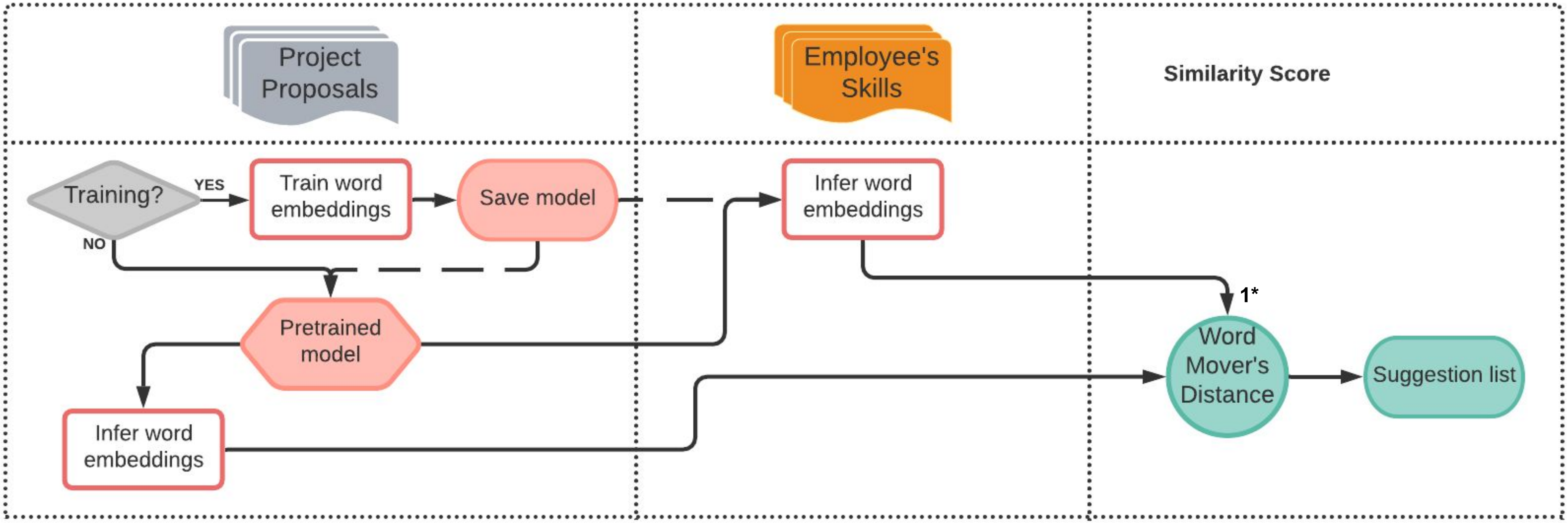
LDA topic modeling and significant words matching



Text embeddings for similarity calculation

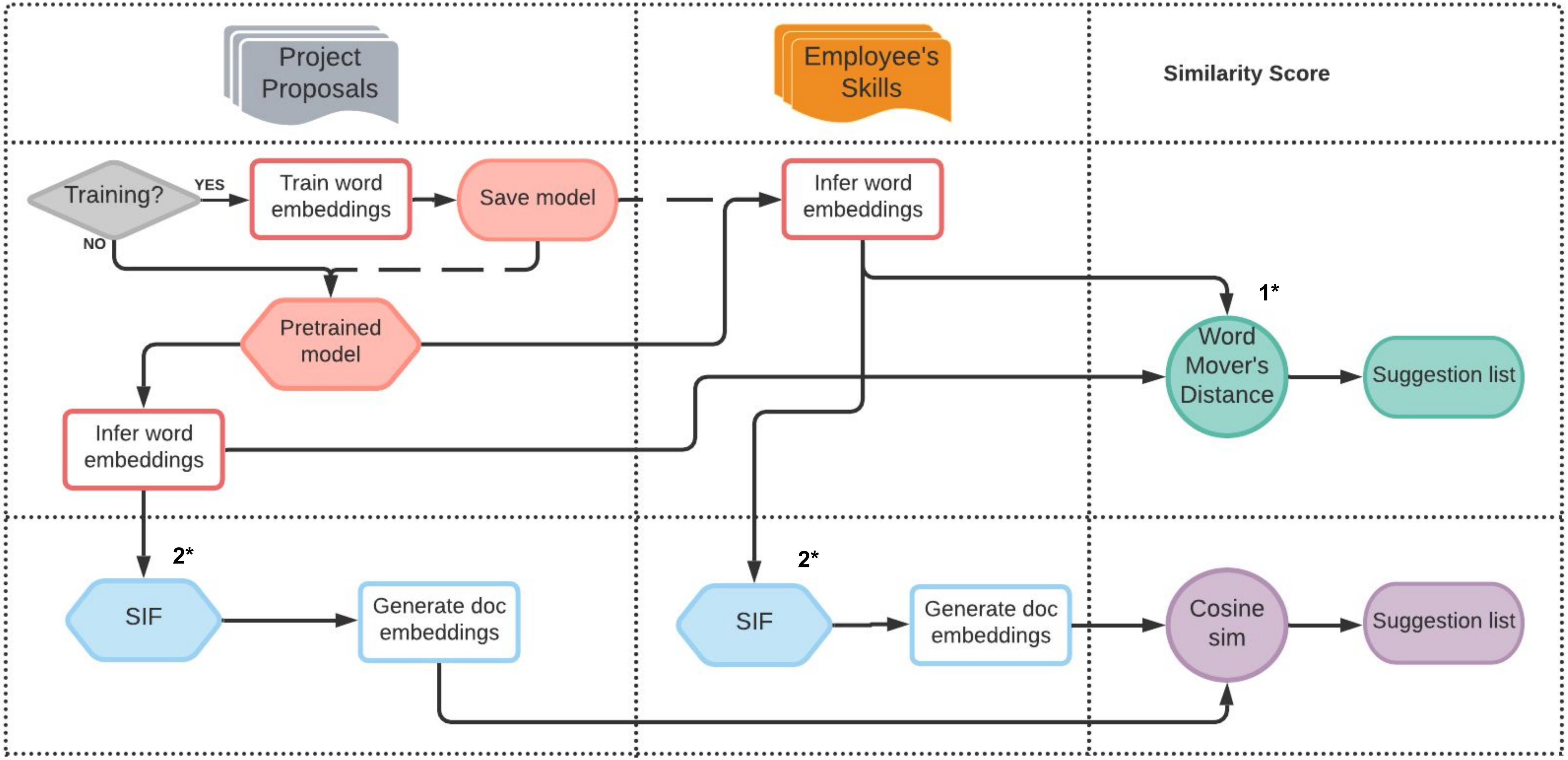


Text embeddings for similarity calculation



Papers: 1*. Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances, 2015

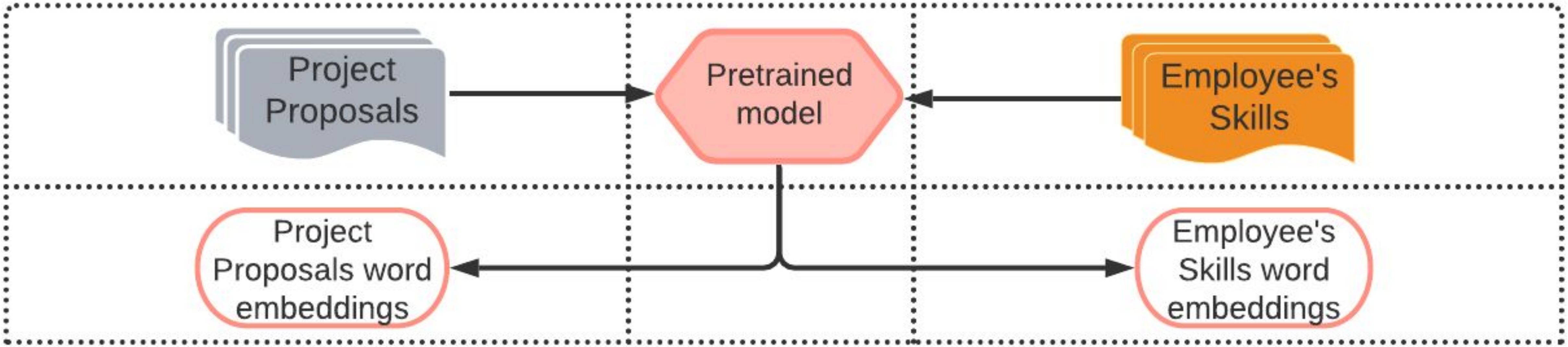
Text embeddings for similarity calculation



Papers: 1*. Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances, 2015

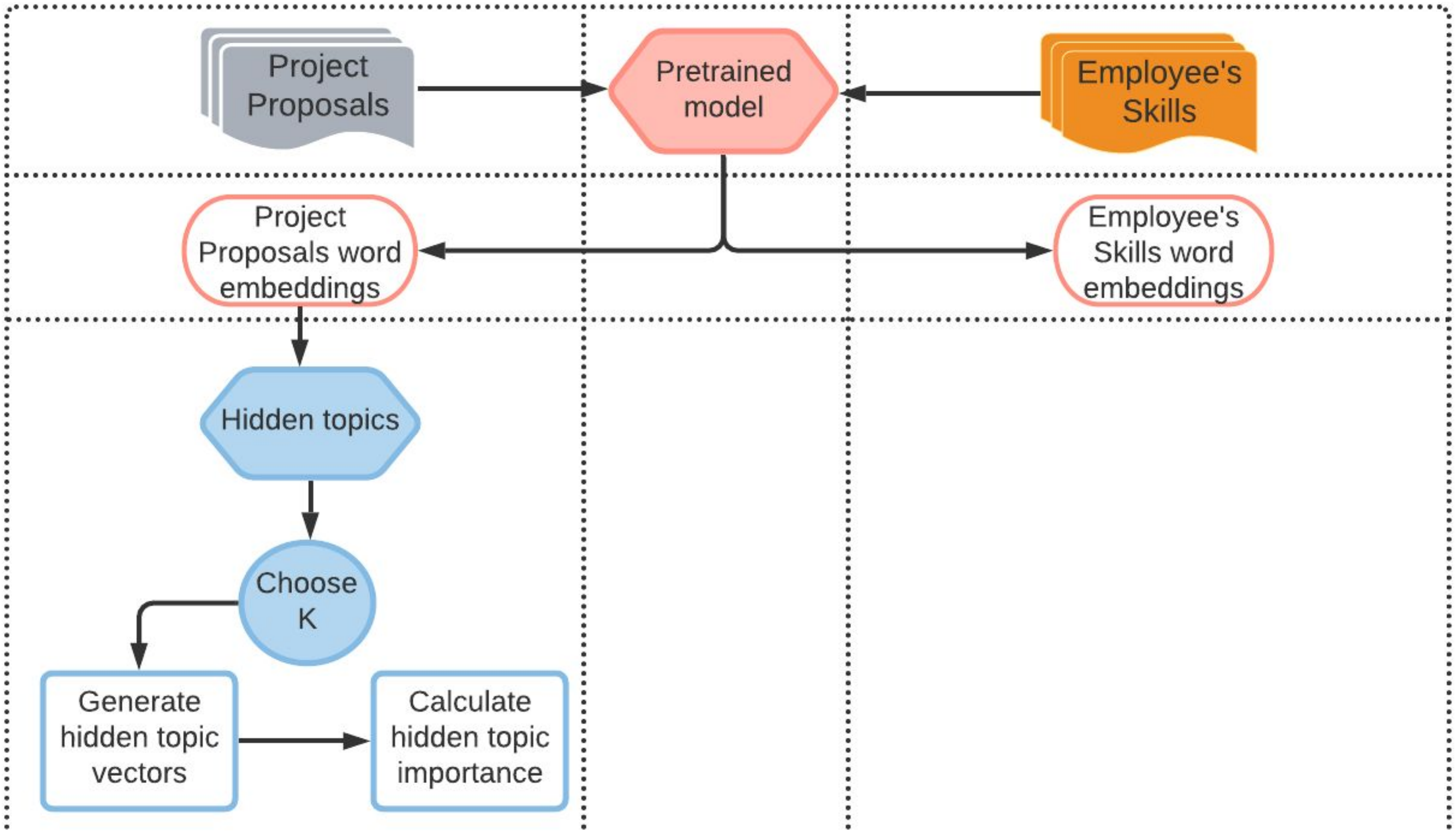
2*. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings, 2017

Matching texts of varying length via hidden topics



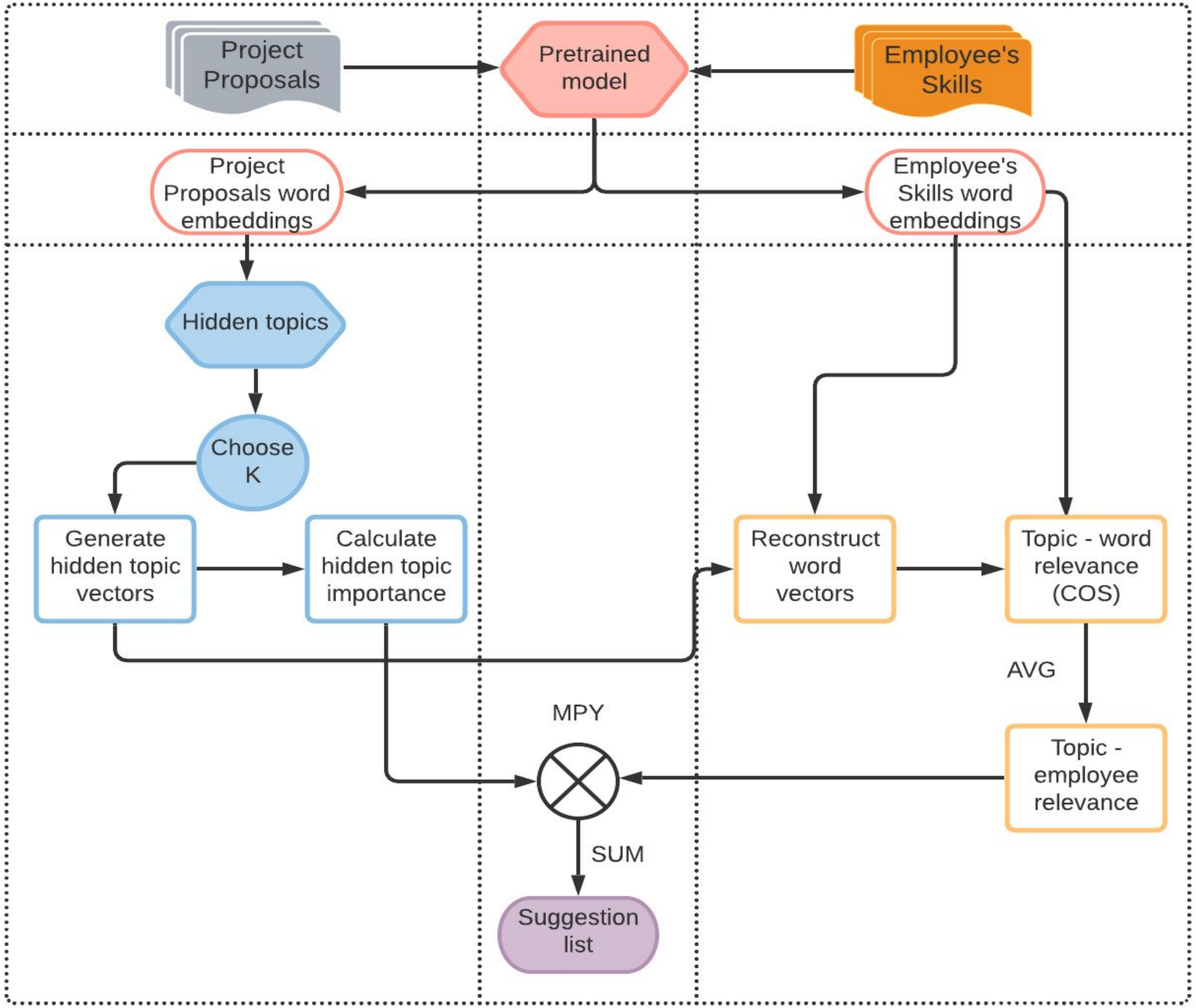
Papers:. Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. Document similarity for texts of varying lengths via hidden topics, 2019

Matching texts of varying length via hidden topics



Papers: Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. Document similarity for texts of varying lengths via hidden topics, 2019

Matching texts of varying length via hidden topics



Paper: Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. Document similarity for texts of varying lengths via hidden topics, 2019

Results

0.5

Model I

LDA topic modeling and significant words matching with number of topics K=90.

0.25

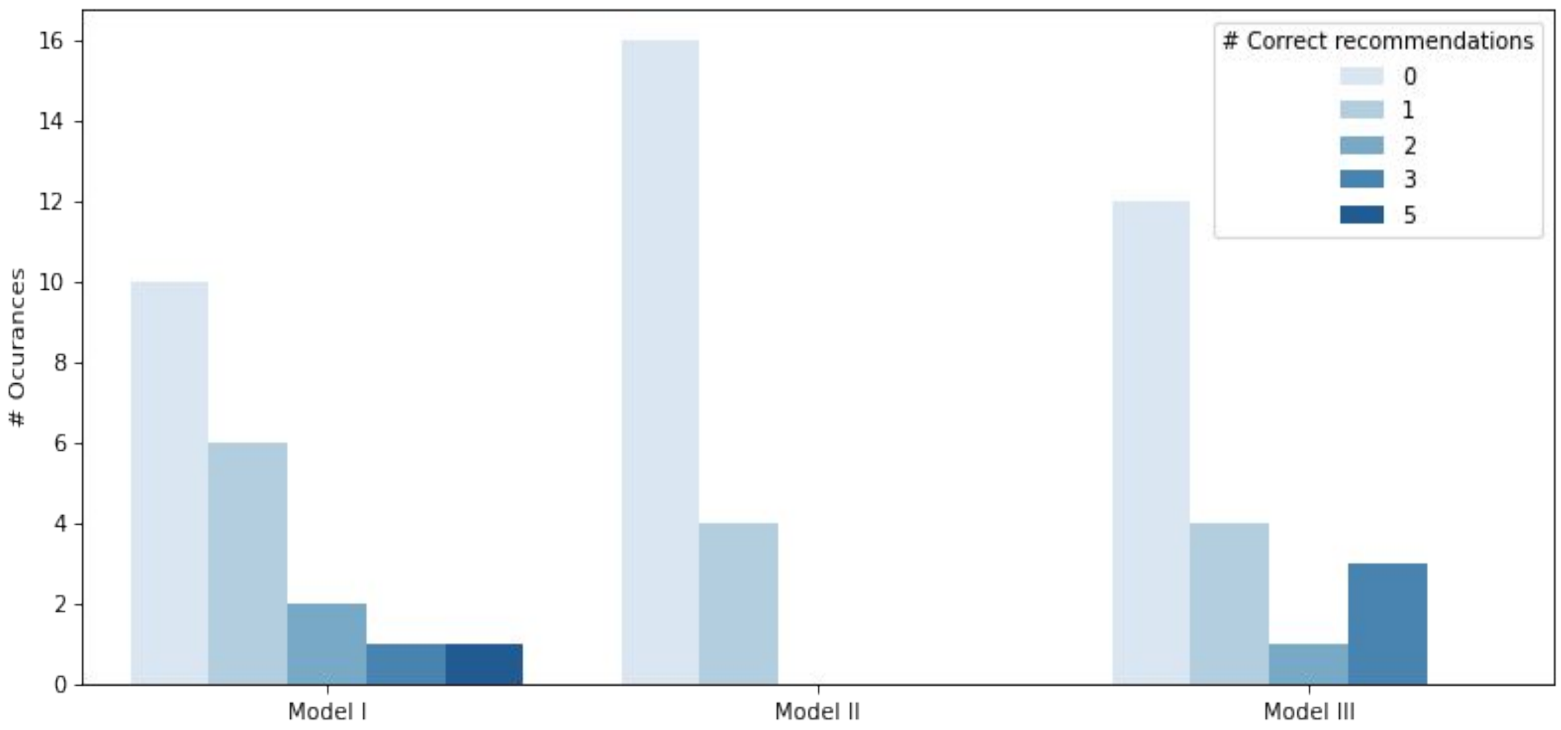
Model II

SIF with self trained word embeddings using fastText model, vector dimension d=300.

0.4

Model III

Hidden topics with pretrained word embeddings trained with fastText model on Common Crawl dataset* , vector dim d=300 and K=5.



*https://fasttext.cc/docs/en/english-vectors.html

Conclusions

01 Text embeddings performed poorly due to the big difference in text lengths.

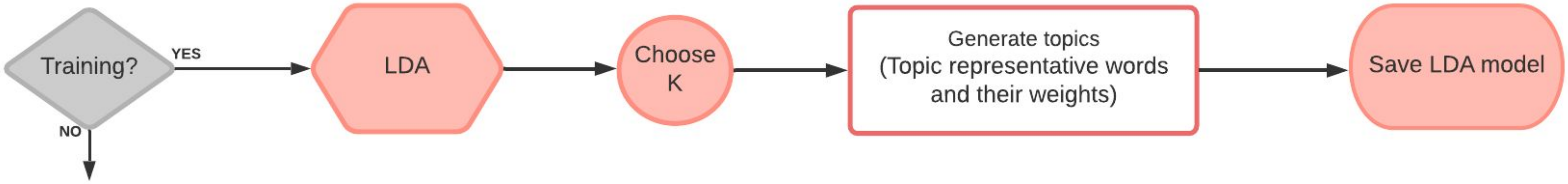
02 Hidden topics approach needs tuning of number of topics K.

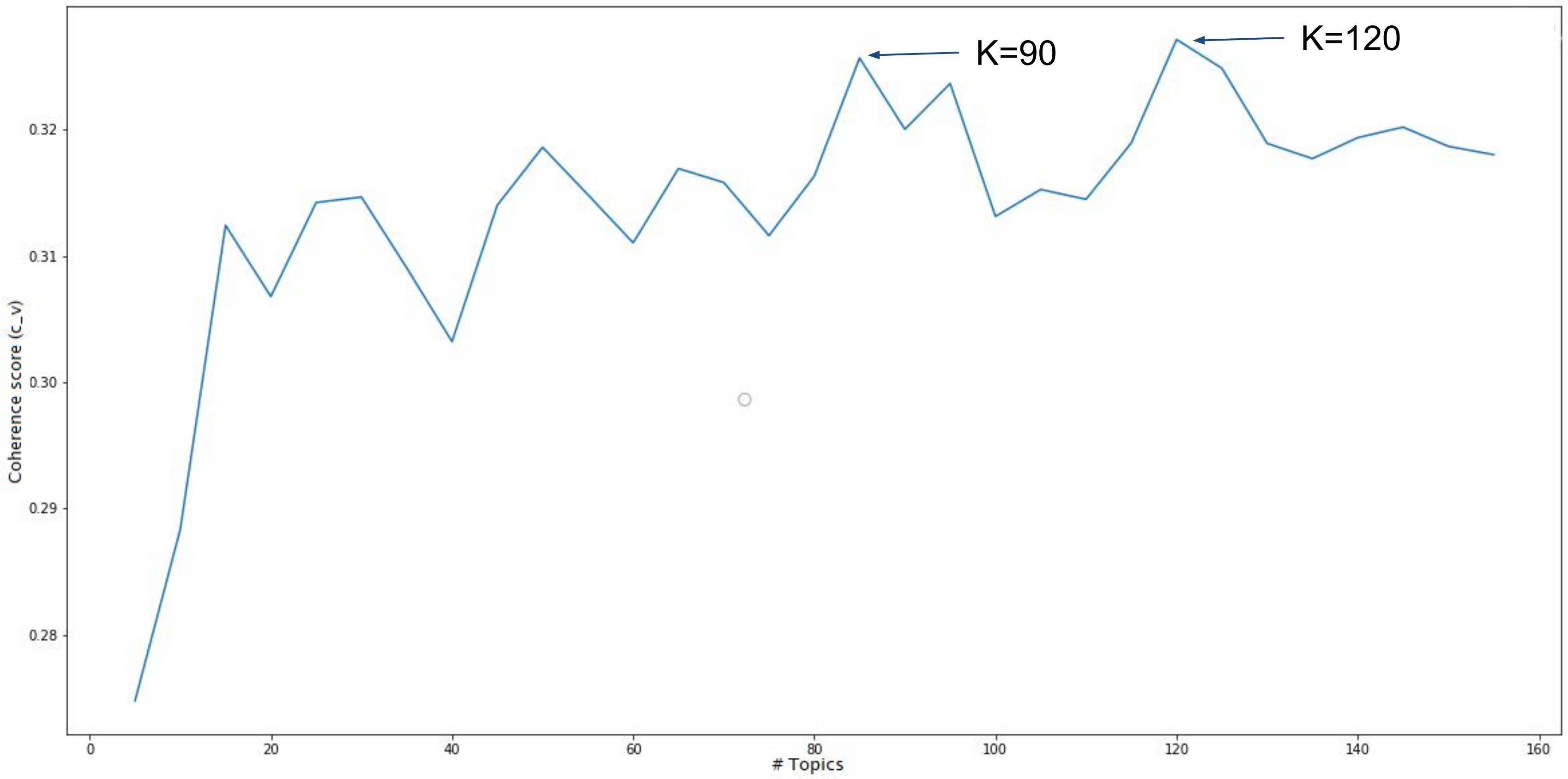
03 The best scored method is: LDA topic modeling and significant words matching.
Further development:

- Full submitted project documentation
- Standardized employee’s skills (ESCO, O*NET)

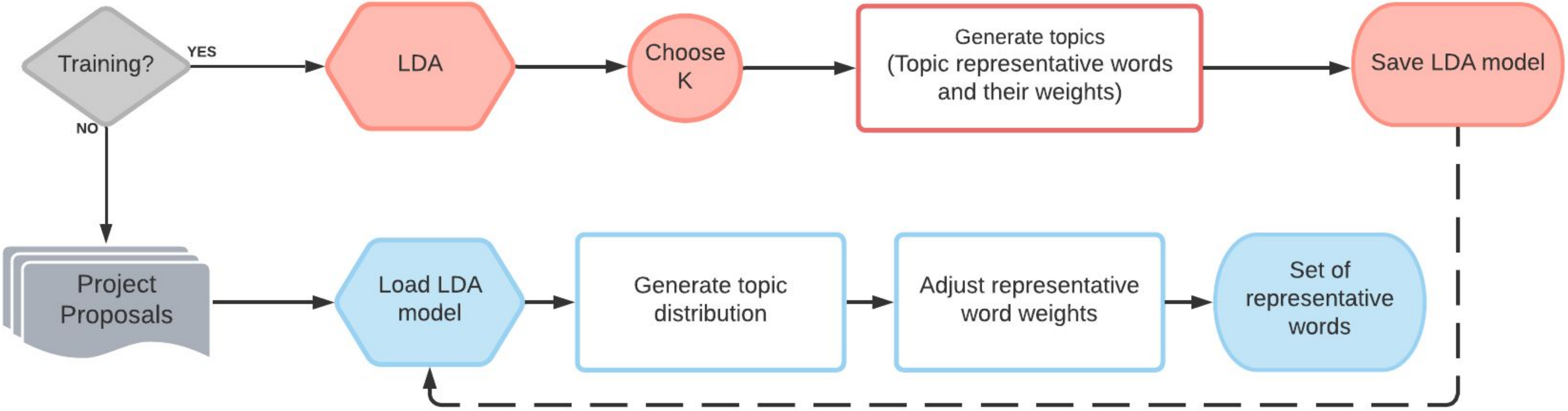
Thank you!

LDA topic modeling and significant words matching

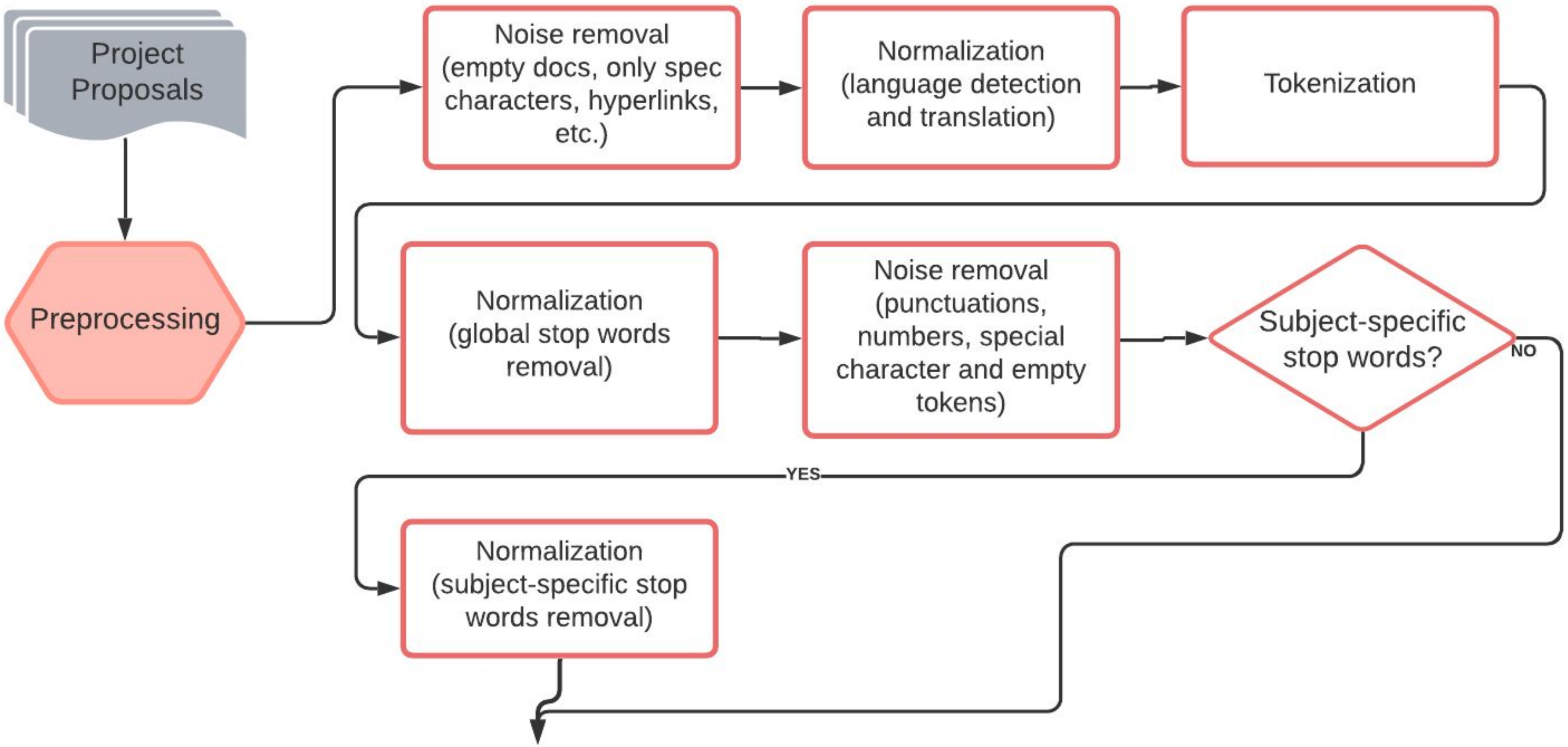




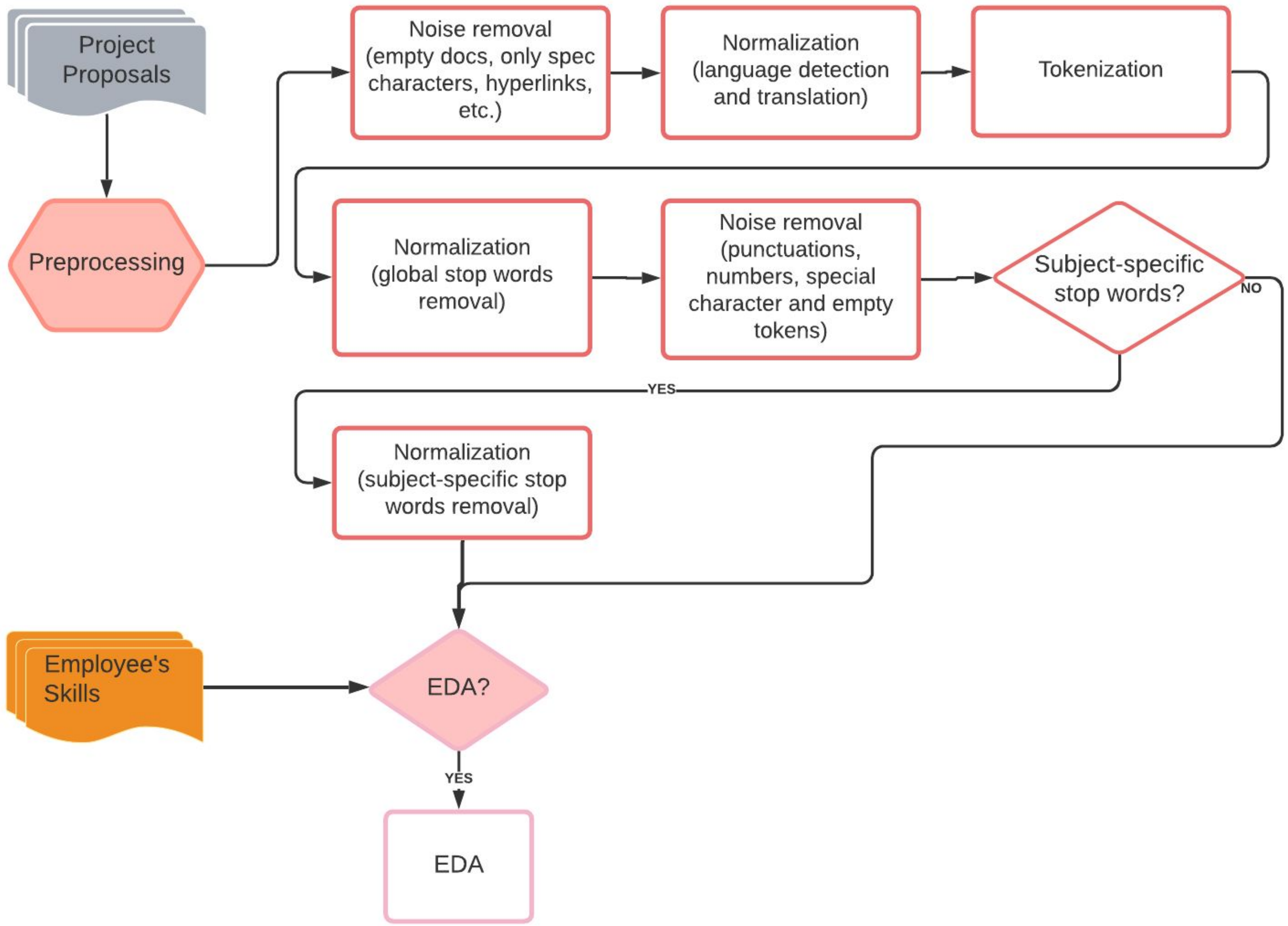
LDA topic modeling and significant words matching



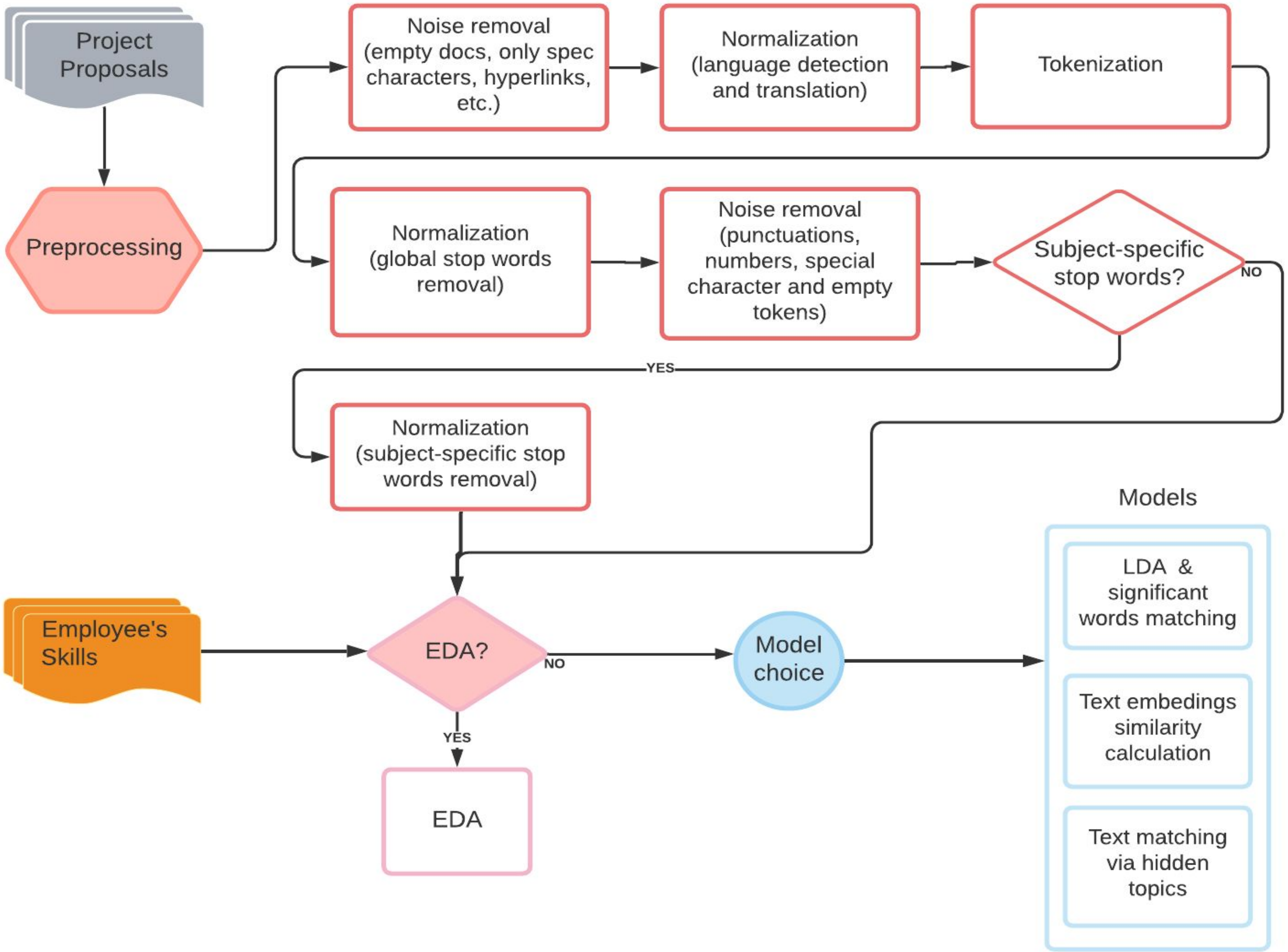
Solution Workflow



Solution Workflow

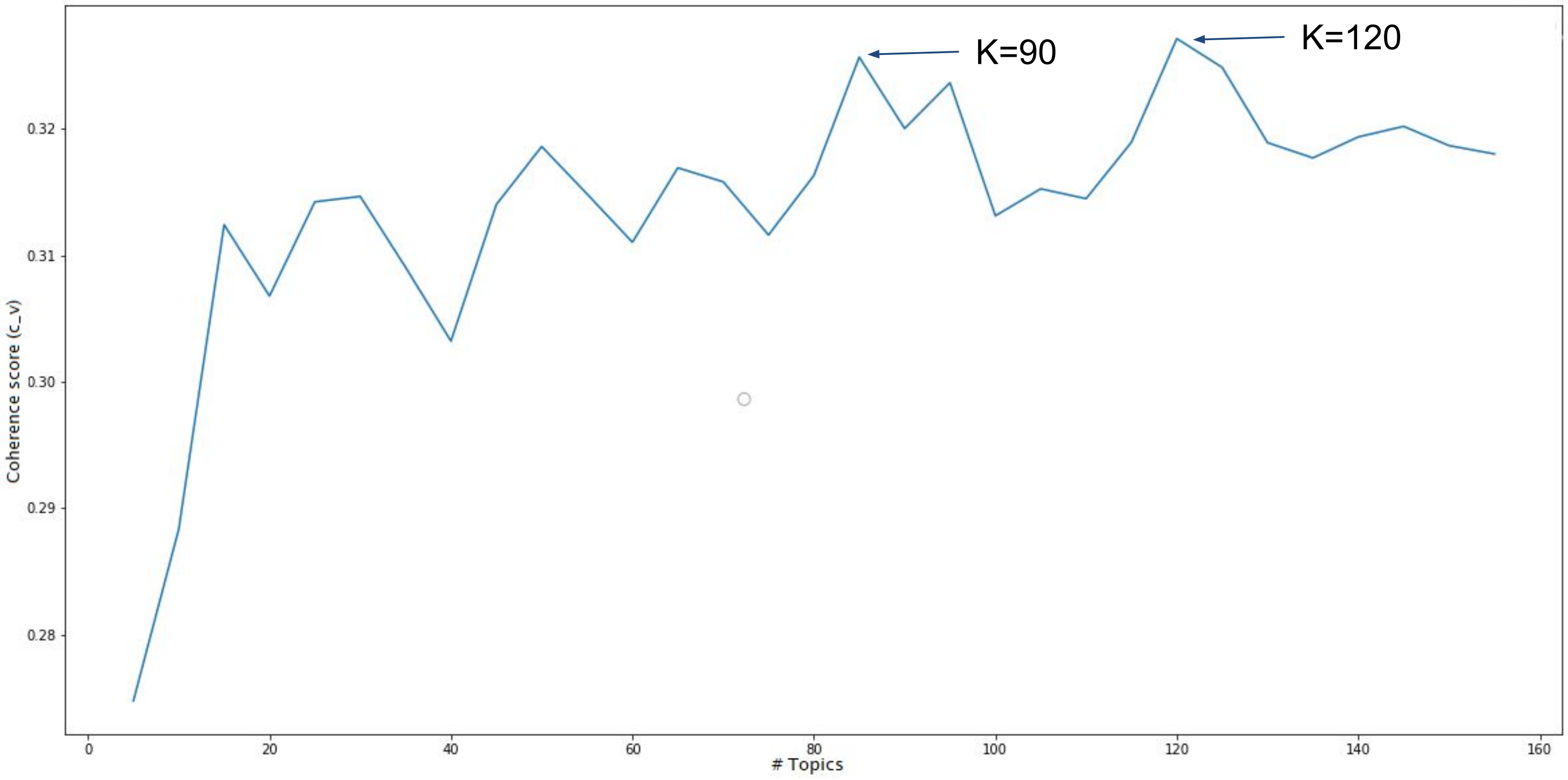


Solution Workflow



LDA topic modeling and significant words matching

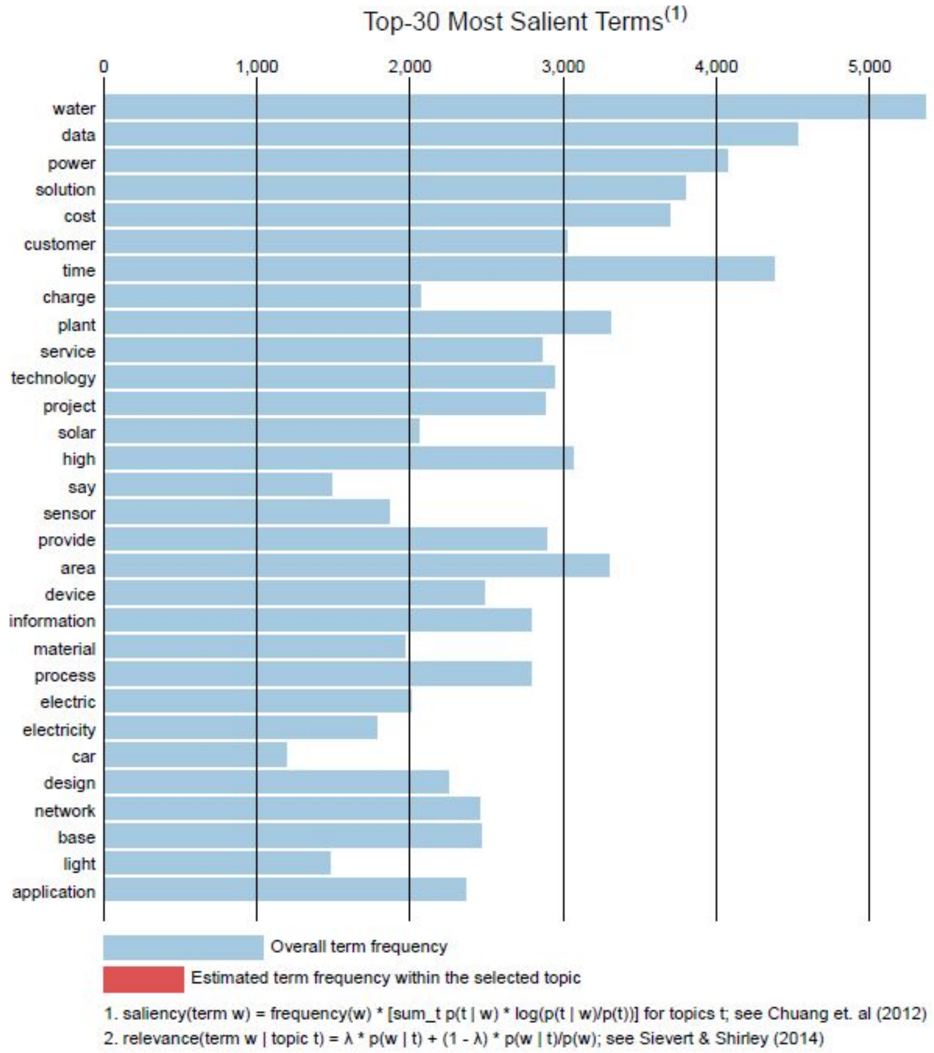
01 CV Coherence score for varying number of topics K.



LDA topic modeling and significant words matching

02

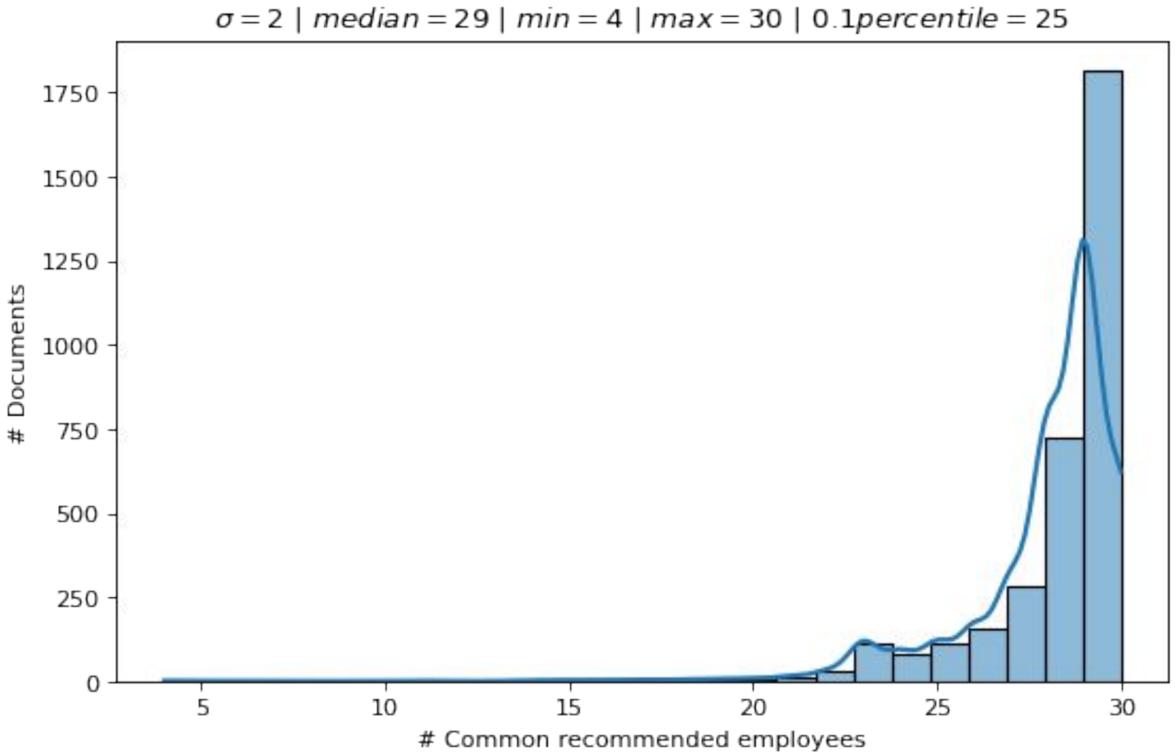
Topics visualisation for K=90.



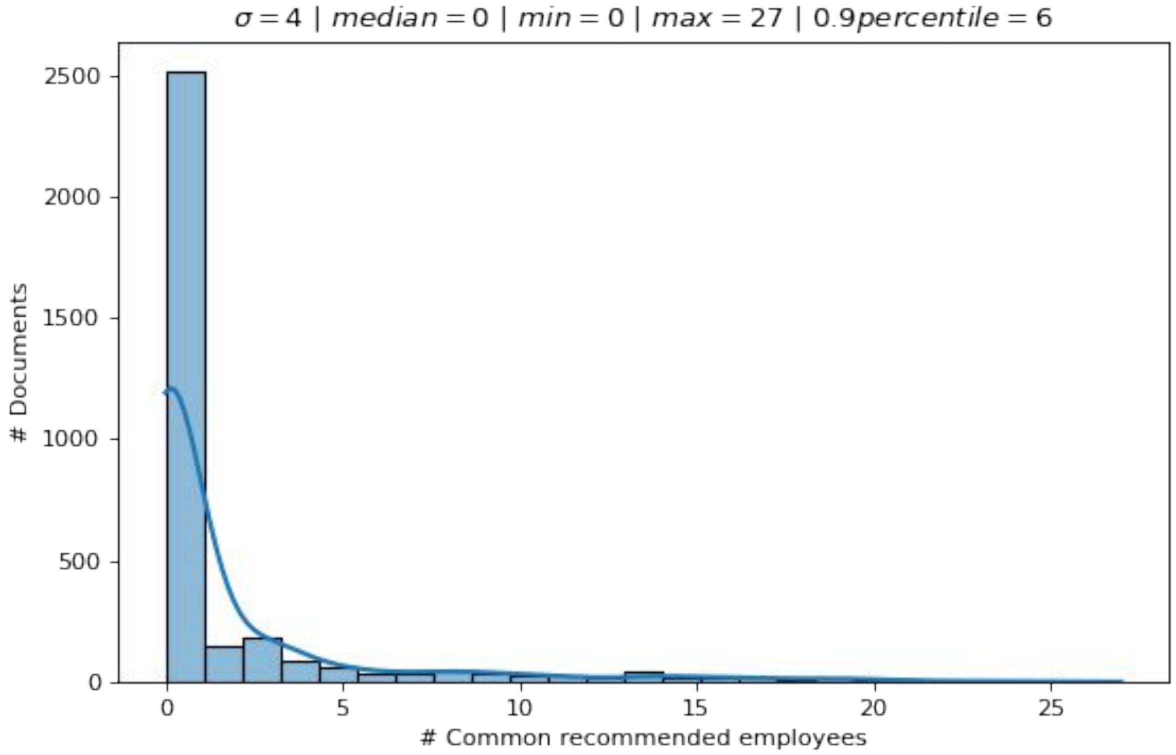
Matching texts of varying length via hidden topics

01

Distribution of number of common employees for K = [5,8,12,18].



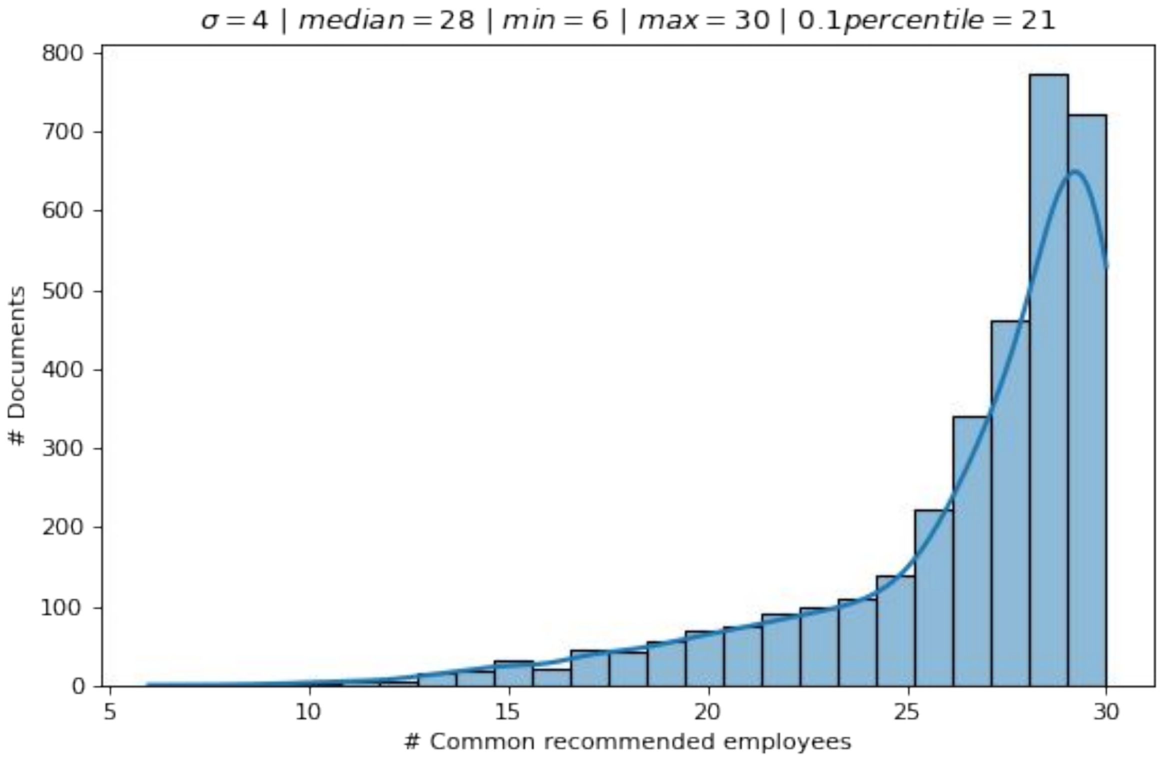
K= 5 & K=8



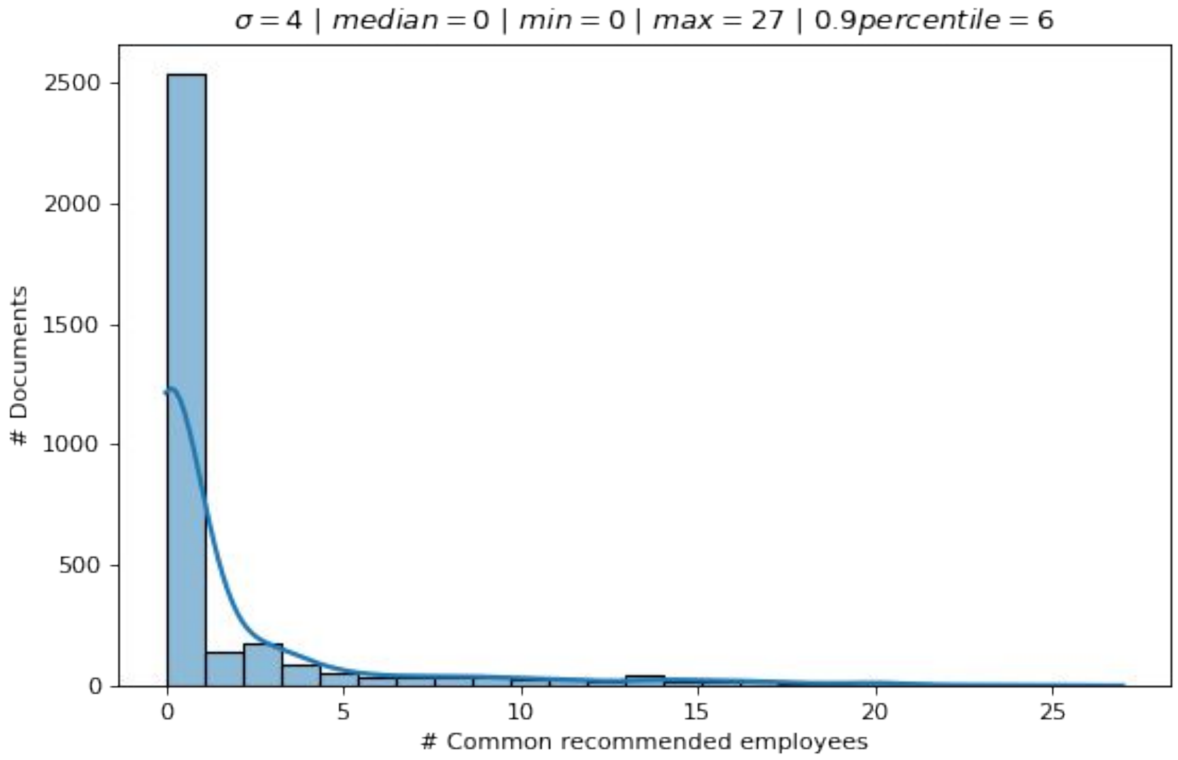
K= 8 & K=12

Matching texts of varying length via hidden topics

01 Distribution of number of common employees for K = [5,8,12,18].

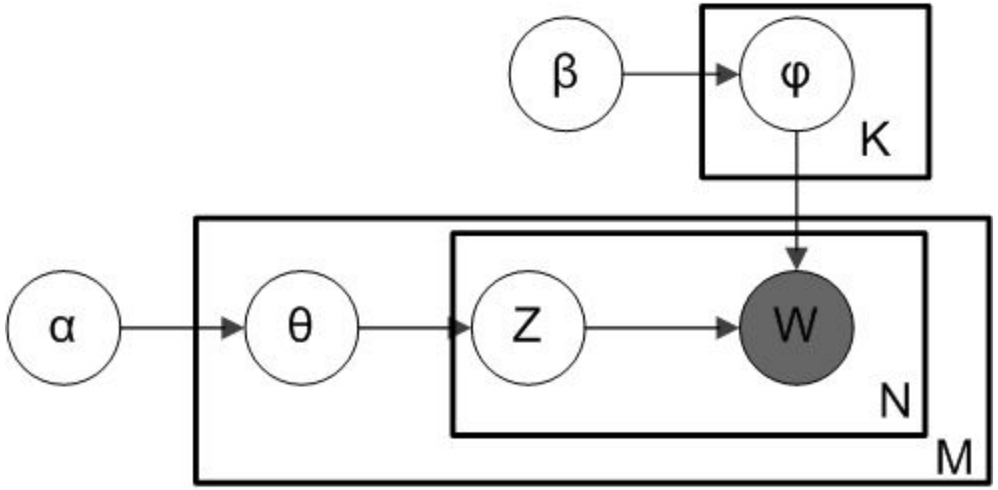


K= 12 & K=18



K= 5 & K=18

LDA



M denotes the number of documents

N is number of words in a given document (document i has N_i words)

α is the parameter of the Dirichlet prior on the per-document topic distributions

β is the parameter of the Dirichlet prior on the per-topic word distribution

θ_i is the topic distribution for document i

φ_k is the word distribution for topic k

z_{ij} is the topic for the j-th word in document i

w_{ij} is the specific word.

3. For each of the word positions i, j, where $i \in \{1, \dots, M\}$, and $j \in \{1, \dots, N_i\}$

1. Cho

(a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.

2. Cho

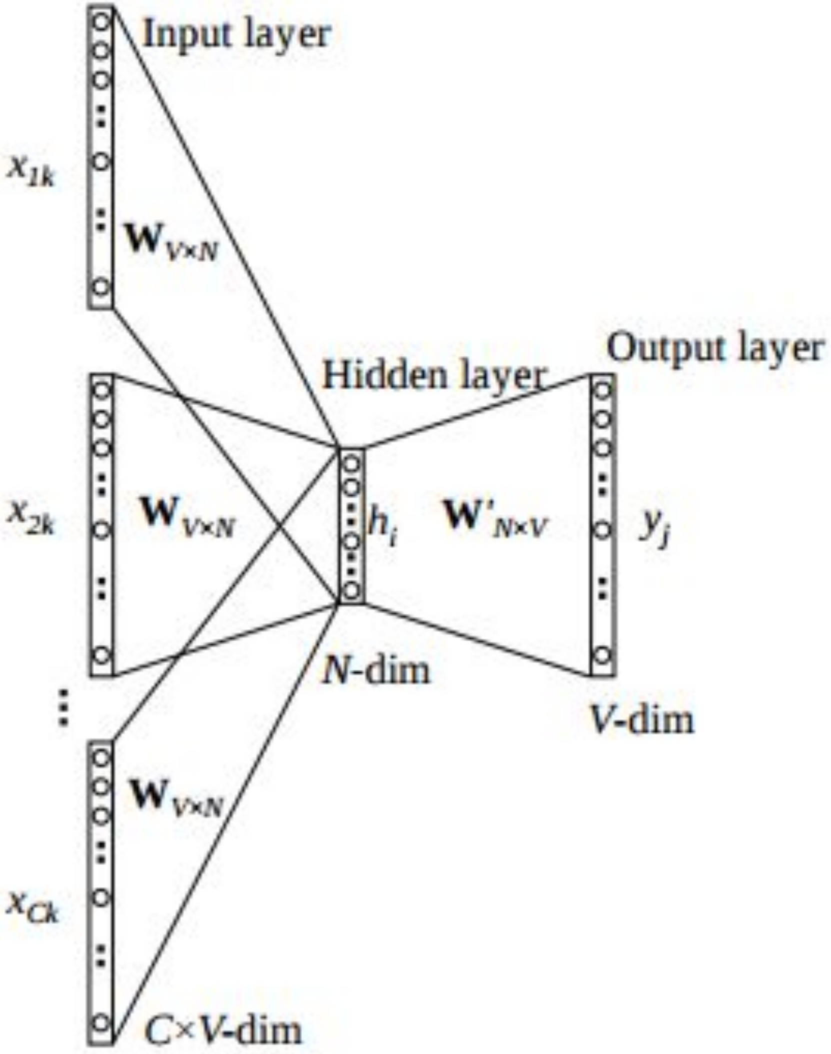
(b) Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

3. For each of the word positions i, j, where $i \in \{1, \dots, M\}$, and $j \in \{1, \dots, N_i\}$

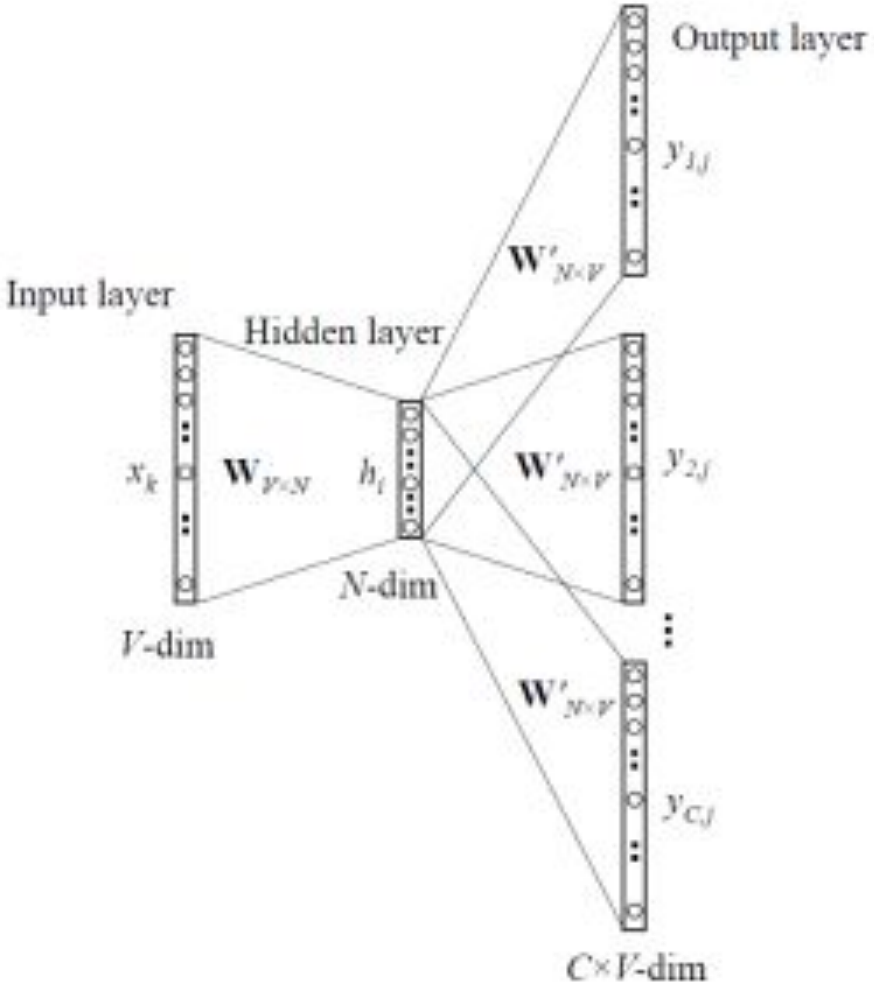
(a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.

(b) Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

Word2Vec



CBOW model



Skip-Gram model

WMD

$$\mathbf{d} = [d_1, d_2, ..., d_n]^T, where$$
$$d_i = \frac{c_i}{\Sigma_j^n c_j},$$
$$c_i = \{word\ i\ appear s\ c_i\ times\ in\ a\ given\ document\}$$

$$c(i,\ j) = ||\mathbf{x}_i\ -\ \mathbf{x}_j||_2$$

$$\sum_j T_{ij} = d_i,$$
$$\sum_i T_{ij} = d'_j$$

$$distance = \min_{\mathbf{T} \geqslant 0} \sum_{i,j=1}^{n} T_{i,j} c(i, j)$$

Matching texts of varying length via hidden topics

$$\min_{\mathbf{H}} \quad \|\mathbf{W} - \mathbf{H}\mathbf{H}^T\mathbf{W}\|_2^2$$
$$\text{s.t.} \quad \mathbf{H}^T\mathbf{H} = \mathbf{I},$$

$$\mathbf{H}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*]$$

$$E_k = \|\mathbf{W} - \mathbf{h}_k^*\mathbf{h}_k^{*T}\mathbf{W}\|_2^2$$

$$i_k = \|\mathbf{h}_k^{*T}\mathbf{W}\|_2^2$$

$$\bar{i}_k = i_k / (\sum_{j=1}^{K} i_j)$$

$$r(\mathbf{h}_k^*, \mathbf{s}_j) = \mathbf{s}_j{}^T \tilde{\mathbf{s}}_j^k / (\|\mathbf{s}_j\|_2 \cdot \|\tilde{\mathbf{s}}_j^k\|_2)$$

$$r(\mathbf{h}_k^*, \mathbf{S}) = \frac{1}{m} \sum_{j=1}^{m} r(\mathbf{h}_k^*, \mathbf{s}_j)$$

$$r(\mathbf{W}, \mathbf{S}) = \sum_{k=1}^{K} \bar{i}_k \cdot r(\mathbf{h}_k^*, \mathbf{S})$$

Exploratory Data Analysis Findings

01 The best grouping of Employee's Skills dataset is on employee level.